



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Retrovirus Evolution

Simon Wain-Hobson

ABSTRACT

Retroviruses manifest a very rich ensemble of genome structures. The evolution of retroviruses varies enormously, with fixation rates varying by as much as a million fold. The emergence of novel genome structures follows remorselessly with the fixation of point mutations and is most apparent for the lentivirus subgroup that has burst on the scene recently. Accordingly, bio-logic suggests that new genome structures will emerge among the lentiviruses, most notably HIV-1.

INTRODUCTION

There is little more ambiguous than words. Yet it is through words that we communicate. Take retrovirus and evolution that make up the title. Once upon a time “retrovirus” was pretty clear and referred to a group of viruses with two copies of polyA⁺ genomic RNA in the virion. Their genetic bauplan was three orfs, *gag-pol-env*, bounded by two long terminal repeats (LTRs). The situation was complicated a little with the identification of

endogenous retroviruses in the mid-1970s and thoroughly disjointed by the discoveries in 1982 and 1983 that both hepatitis B virus (HBV) and cauliflower mosaic virus (CaMV) replicate by way of an RNA intermediate. The infectious form of both of these viruses is DNA. However, given that the history of “retroviruses” goes back to the early 1900s, that they induced cancer, while the remarkable phenomenon of reverse transcription that broke “The Central Dogma” of the day, i.e. DNA makes RNA makes protein, was established using these oncogenic retroviruses, classical “retrovirology” understandably had the advantage of terrain.

HBV was referred to as a “pararetrovirus” before becoming the prototype of the “hepadnavirus” group. CaMV lent its name to the caulimovirus group. Retrovirology has come to be synonymous with classical retroviruses.

Yet just at this moment, retrovirology was about to be discombobulated once again, and this time by an insider. Prior to HIV, the key retroviruses were oncogenic, otherwise known as the avian and murine leukemia viruses (ALV and MLV) and mouse mammary tumor virus (MMTV). The discovery in 1980 of human T cell leukemia virus (HTLV) associated with

adult T cell leukemia virus fitted the bill, so all was well. However, with the isolation of HIV-1, a lentivirus, and a major new disease, AIDS, indeed the infection of a generation, everything changed. It coincided with an unprecedented firepower in molecular biology that has led us to generate almost more data than we know what to do with. Within 2 years of isolation detection kits became available, with the first designer antivirals appearing by 1995. HIV is the number 1 virus for the *Journal of Virology*. The younger generations of students know little of ALV, MLV, and MMTV. As ever such a magnet can warp vision, and anything and everything new is tried and tested on HIV, a little like the involvement of every new virus in multiple sclerosis (Vandvik and Norrby, 1989; Winkelmann, 1993; Ablashi *et al.*, 2000).

“Evolution” has almost as many meanings as researchers working on it. Outside of a biological setting, evolution actually refers to dy/dx , the change of one parameter with respect to another. Of course in biology we are interested in the temporal component, hence it is dy/dt that is important. Yet for Darwin the “stuff of evolution” was adaptation, in fact repeated adaptation that over time led to something as wondrously complex as the eye probably 40 times or so.

Coming to the point, this chapter will deal with the evolution of the classical retroviruses. Evolution will be considered from the macroscopic point of view which, considering that genomics dominates contemporary microbiology, is tantamount to the generation of novel genomic organizations. And this is where we will start.

GENOME ORGANIZATION AS A STARTING POINT

Retroviruses are enveloped viruses composed of a lipid bilayer surrounding a nucleocapsid within which resides the two copies of genomic plus stranded genomic RNA. Understandably the genome must code for capsid and envelope proteins. Given the provirus,

the integrated DNA replication intermediate, a reverse transcriptase and integrase are called for. It turns out that three open reading frames (orfs) are sufficient to generate a fully functional retrovirus. Retroviral genetic maps are invariably given in the proviral form because it is that of the gene (Figure 13.1). With the signal exception of the spumavirus group of retroviruses, the provirus is a single, multiply spliced gene. The words gene and orf are used interchangeably in retrovirology which is, of course, an abuse of language. The convention is that genes/orfs are referred to in italics while the corresponding proteins are given in normal type. The *gag* orf encodes the nucleocapsid proteins in the form of a polyprotein precursor. *gag* is followed by *pol*, encoding a large polyprotein incorporating the mature reverse transcriptase; RNase H and integrase proteins. Maturation of the polyproteins is the accomplishment of the viral protease that is generally part of *pol*, but can exist in a separate orf termed *pro* located between *gag* and *pol*. Uniquely for Rous sarcoma virus (RSV) is the protease attached to the C-terminus of the Gag polyprotein. The third orf is *env*, and encodes a polyprotein envelope precursor that is matured by host cell proteases. For ALV, MLV, and MMTV, the historical benchmark retroviruses, *gag* is always followed by *pol* which is always followed by *env*, and *gag-pol-env* represents the lowest common denominator for retrovirus genome organization.

However, as can be seen from Figure 13.1, this effective yet minimalist view of retroviral genomics is far from common. The most complex retroviruses bear up to nine distinct orfs encoding 15 mature proteins. Some orfs are unique to a virus, others common to a cluster. Some orfs are apparently analogues even though there is no sequence identity between them. While the *gag-pol-env* synteny is never violated, a multitude of orfs can separate *pol* and *env*. Frequently *env* is followed by 1–4 orfs, although an orf preceding *gag* is rare.

Such exuberance in terms of genetic organization is probably without precedent among small viral genomes, although coronavirus genome organization shows considerable

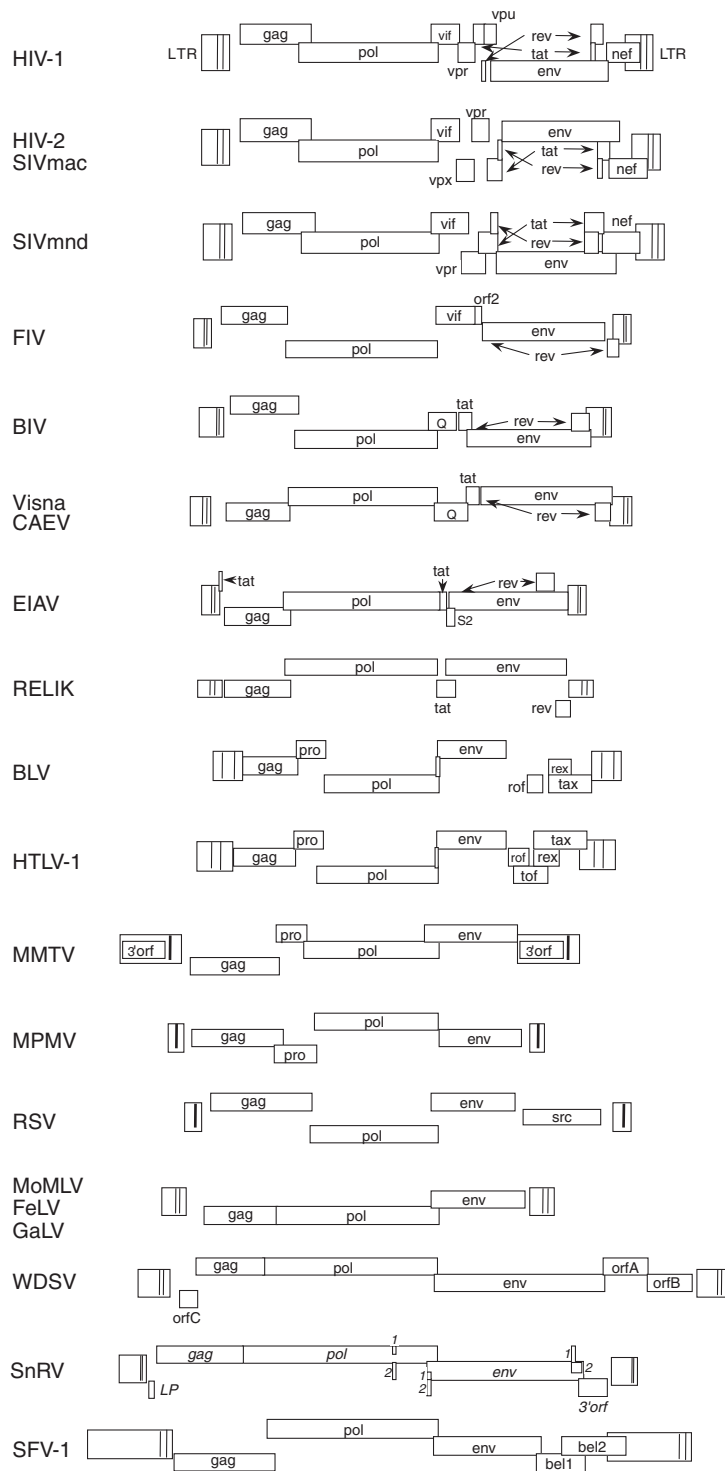


FIGURE 13.1 Retroviral genome configurations. These are given in the DNA, or proviral, form found integrated into host chromosomal DNA. The open reading frames are represented by boxes flanked by the LTR (long terminal repeat), which encodes most of the information necessary for transcription and reverse transcription. In some cases there is considerable amino acid identity, as in the case of *pol*. However, although *env* indicates the envelope orf there is no sequence identity whatsoever between, say, those of HIV, HTLV-1, RSV, and WDSV. The same is true of the retroviral transactivators *tat*, *tax*, and *orf1*. Even for a given virus, strain variation can attain staggering proportions (see Figure 13.5).

plasticity. Of course herpesviruses genomes span 119–236 kb while poxvirus genomes vary even more—134–360 kb. By contrast the three-fold range in the number of retroviral orfs is accompanied by a relatively modest ~50% variation in the size of the coding region.

Where do all these “extra” genes come from? The origin of the RSV *src* gene is no secret, it being transduced from the avian genome (Stehelin *et al.*, 1976), while the primate lentiviral *vpx* gene is arguably derived from a duplication of the *vpr* gene (Tristem *et al.*, 1990). For the rest, sequence homology is non-existent. It is possible that many are of cellular origin, however, given the tempo of retroviral mutation (*vide infra*), all traces have been lost.

STABILITY OF GENE ORGANIZATION YET AT LEAST SIX WAYS TO EXPRESS POL

Apart from gene transduction, the genomic organization of a particular retrovirus is invariant. While it is possible to isolate viable strains carrying deletions say in *env* (JDV) or *nef* (HIV, SIV), deletions are common fare among microorganisms. As the provirus represents a single gene, large rearrangements and inversions are presumably deleterious. When making retrovirus-wide comparisons of the common *gag*, *pol*, and *env* orfs, the internal order of the major proteins is invariably conserved. Thus for Gag the order is always matrix antigen, capsid antigen, and nucleic acid binding protein, for Pol the order is protease, reverse transcriptase/RNaseH, integrase while for Env, the order is surface protein, transmembrane protein.

Nonetheless, there are a couple of noteworthy cases of reorganization. The first concerns the protease. As can be seen, it always precedes the reverse transcriptase/RNaseH (Figure 13.2A). However, the coding sequences can be part of *gag*, *pol* or in a separate reading frame, called *pro*. In view of this it is not surprising to learn that there are several distinct mechanisms by which these distinct

reading frames are connected. The mechanisms range from –1 ribosomal frameshifting once, twice, splicing, and suppression of a terminator codon. If the larger family of retroelements are considered for once, there are two additional Gag-Pol syntenies involving +1 ribosomal frameshifting and the two being part of one long polyprotein precursor. These examples show that while the gene/orf order may be preserved, there are radical differences underlying their expression.

The second example concerns the retroviral dUTPase (McGeoch, 1990). Interestingly, only a subset of lentiviruses and the MMTV/SRV group encode such an enzyme and then in two different locations (Figure 13.2B). For the lentiviruses, it fits snugly in phase between the RT/RNaseH and integrase coding regions, while for the MMTV/SRV group it is located upstream of the viral protease. While phylogenetic analysis shows that they cluster together, they do not constitute a robust monophyletic group, meaning that it is not possible to distinguish between a single introduction or independent introductions into each group (McGeoch, 1990). This is the only example of radical change in the synteny of retroviral coding sequences. Of course for large DNA viruses, such as the mycophages, marine phaeoviruses, and herpesvirus, to name but three, extensively rearranged genomes are commonplace (Weir, 1998; Delaroque *et al.*, 2003; Pedulla *et al.*, 2003).

ERROR AND RECOMBINATION

When it comes to making replication errors, DNA and RNA virus differ hugely. RNA viruses are unable to undertake proofreading or mismatch repair while DNA viruses can, either because DNA replication is performed by the host cell or else the genome is large enough to encode the necessary enzymes to undertake high-fidelity replication. In this latter context, poxviruses replicating in the cytoplasm is perhaps the most extreme example. Where do retroviruses, half DNA, half RNA viruses sit when it comes to error? Errors

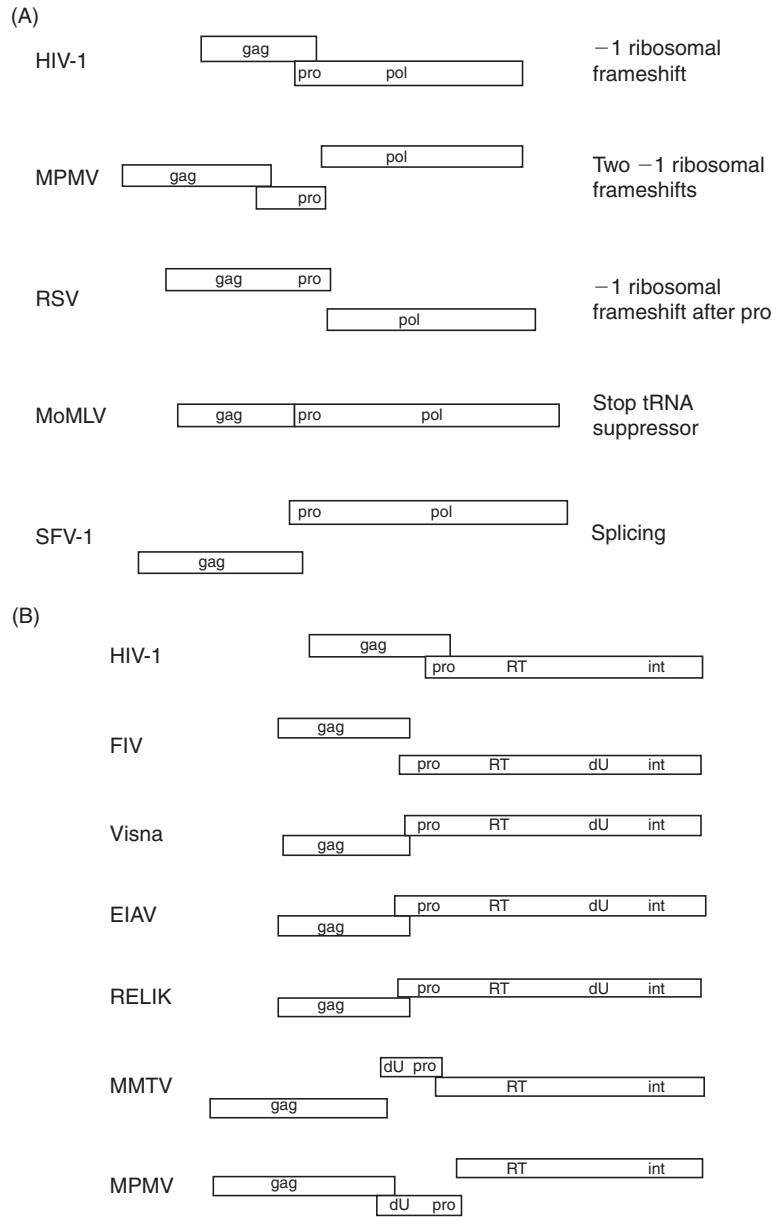


FIGURE 13.2 Variable synteny for the retroviral protease (*pro*) and dUTPase (*dU*) coding sequences. (A) Not only is the *pro* coding region in different reading frames, the mechanism by which it is expressed in the form of a Gag-Pol polypeptide precursor is radically different. (B) The dUTPase coding sequence is found only in a subset of retroviral genomes and then not consistently. For example, all the primate lentiviruses do not carry a dUTPase.

made during reverse transcription per se are not corrected whatsoever, while those made during plus-strand synthesis could be proof-read when the DNA is translocated to the

nucleus. It turns out that although retroviral mutation rates are in the range of 0.3–0.05 per genome per cycle (Mansky and Temin, 1995; Mansky and Wisniewski, 1998; Mansky, 2000),

they are approximately an order of magnitude lower than those for RNA viruses, 1–3 per genome per cycle (Drake and Holland, 1999). While it would be tempting to ascribe this difference to proofreading of mutations of the plus-strand errors, as more errors are made during minus-strand synthesis, it probably reflects some subtle intrinsic difference between the polymerases.

As in almost all polymerases, transitions are produced more readily than transversions, while deletions are more frequent than insertions (essentially duplications). For the HIV reverse transcriptase and presumably the lentiviruses as a whole, deletions arise more frequently than for other RTases. Being diploid viruses, recombination is to be reckoned with. Retroviral recombination goes back to 1971 (Vogt, 1971; Kawai and Hanafusa, 1972; Weiss *et al.*, 1973) just one year following the description of reverse transcription. It occurs essentially by a copy choice mechanism (Coffin, 1979). Precise recombination rates have been worked out for two retroviruses, HIV-1 and MLV, the values being 3–4 per virus per cycle respectively for both viruses, despite an initially lower rate for MLV (Jetzt *et al.*, 2000; Zhuang *et al.*, 2006). Importantly, these values are $\sim 10\text{--}100\times$ the point mutation rate for both retroviruses. This means that as soon as a point mutation is made it will be recombined in the following round.

There are a couple of conditions to be met before concluding that everything is recombined for a retrovirus. First, infected cells have to be multiply infected, meaning that they must harbor multiple proviruses. Second, the proviruses must be genetically distinct. If, for example, a multiple infected cell received all its HIV content from a productively infected neighboring cell that harbored a single provirus, then recombination would proceed but effectively produce the parental virus despite recombination. So does multi-infection occur and are the proviruses divergent *in vivo*?

We have only the beginnings of an answer to these questions and then only for HIV-1. Using fluorescent *in situ* hybridization (FISH) a single report showed that for two patients, one with high and one with low plasma viremia, the average proviral copy number

in splenic mononuclear cells was 3–4 with a range of 1–8 (Jung *et al.*, 2002). The majority of cells ($\sim 72\%$) were multiply infected. Laser microdissection of the FISH positive nuclei and transfer to PCR tubes allowed amplification and sequencing of the HIV-1 DNA. The sequence complexity of the V1V2 hyper-variable regions analyzed, probably the most sensitive indicator of Env variation, was stunning. Up to 28% amino acid variation was found within a single nucleus. This is greater than the average protein variation between humans and birds! To add to a complex situation, numerous distinct recombinants were also found within the same nucleus.

What does this mean? First, it indicates that phenotypic mixing is possible, generating heterogeneous virions (Figure 13.3B). Second, it means that HIV virions are non-clonal objects. Worse, these non-clonal structures are ephemeral as production, infection and reverse transcription involving recombination are a matter of 10–18 hours (Figure 13.3). Such complexity inside a single nucleus (synonymous here with boundary or discrete volume) indeed conjures up the quasispecies. Indeed, is this the true quasispecies where genomes and gene products clearly interact? The population of genomes is of course limited to two or at most eight functional genomes, which can produce a rich variety of heterokaryons. However, they will recombine in the next round of infection producing a new ensemble of recombinant genomes. This leads to the third point. How is fitness maintained? Perhaps this question should be rephrased into a more prosaic form, what fraction of recombinants is neutral and what fraction deleterious? The concise answer is we simply don't know. However, when recombinants of HIV and SIV are made the resulting "SHIV" that is eventually recovered invariably carries "additional" substitutions as though the jump in sequence space made by the molecular virologist landed the SHIV in a relatively unfit region. It seems possible that most recombinants may in fact reduce fitness. Certainly, it seems logical given that functional sequence space can represent only a small fraction of total sequence space. Could the retroviral

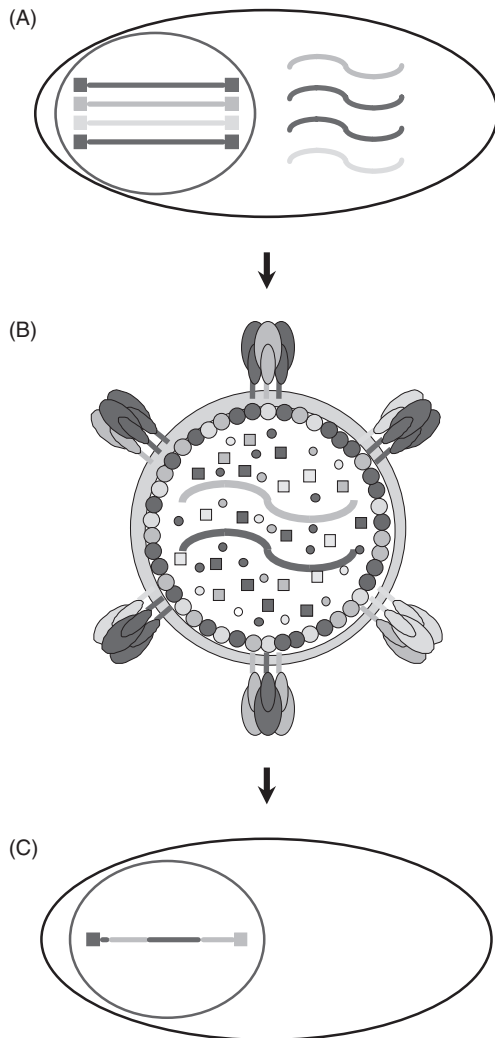


FIGURE 13.3 The non-clonal and ephemeral nature of HIV. (A) A cell bearing four genetically distinct proviruses is represented by different shading. (B) Upon productive infection, only two RNAs can be packaged per virion and there is no reason to consider that proteins from all four proviruses are incorporated into the virion. The resulting structure is much like a harlequin. No two virions will be the same. It is not known if this impacts the biology of the virus. (C) Upon infection of the target cell, recombination occurs leaving a mosaic provirus. All this occurs in as little as 18 hours. (See Plate 15 for the color version of this figure.)

Red Queen be constantly running to make up for the devastating effects of recombination? Could this be the major driving force leading to the relentless accumulation of point mutations

for HIV, no matter the multiplicity of infection, the disease stage, or the nature of existing immune responses?

At the more macroscopic level, recombinants are found at all levels of HIV and SIV—intrapatent recombinants, those among different strains of the same clade, the ever-expanding number of circulating recombinant forms (CRFs) worldwide that can be readily shown to be mixes of the original founder clade viruses, recombinants of CRFs, inter-group recombinants, e.g. HIV-1 M-O (Peeters *et al.*, 1999), and a plethora of recombinants between HIV-1 precursor viruses with SIVs of chimpanzees (Bibollet-Ruche *et al.*, 2004). Many of the diverse SIVs from small equatorial monkeys are demonstrably recombinants (Jin *et al.*, 1994). Indeed we are probably at a level where the null hypothesis should be that everything is recombinant unless shown otherwise.

One of the hottest topics in retrovirus variation right now is APOBEC3-mediated hypermutation, which is extensively treated in the excellent chapter by Warner Greene (Chapter 8). Briefly, these are cellular cytidine deaminases that if packaged into the virion can edit hundreds of cytidine residues, leaving behind uracil. This is tantamount to a host restriction system. It is particularly a problem for the lentiviruses to the point that they have evolved the *vif* gene that prevents APOBEC3 incorporation (Sheehy *et al.*, 2002). While other retroviruses including HBV do undergo editing *in vitro* and *in vivo* (Mahieux *et al.*, 2005; Suspene *et al.*, 2005; Delebecque *et al.*, 2006), there is no evidence that this seriously impacts their replication. For the mouse there is a single APOBEC3 gene and knockout mice are both viable and fertile, which suggests that it does not restrict the transposition of endogenous retroviruses (Mikl *et al.*, 2005). This fits in well with the fact that although the bird genome carries a plethora of endogenous retroviruses, there is no APOBEC3 counterpart (Harris and Liddament, 2004). Might a little APOBEC3 editing help the lentiviruses evolve? Could it occasionally act as the equivalent of a bacterial mutator? As is frequently the case, speculation is understandably inversely proportional to the amount of data.

RETROVIRAL FIXATION RATES VARY HUGELY

It comes as no surprise to any virologist that fixation rates vary with the protein under study—the flu paradigm is known to all. To get this quickly out of the way, the retroviral surface envelope protein is generally the most variable, although some of the small accessory proteins fix mutations at a comparable

rate. The nucleocapsid Gag proteins are understandably less tolerant of change only in the temporal sense, while the enzymes that make up Pol bring up the rear. The retroviral integrase has the lowest amino acid fixation rate and so it not a surprise that this region is invariably used to perform retrovirus-wide phylogenetic analyses. A typical phylogenetic tree based on the integrase gene is shown in Figure 13.4A.

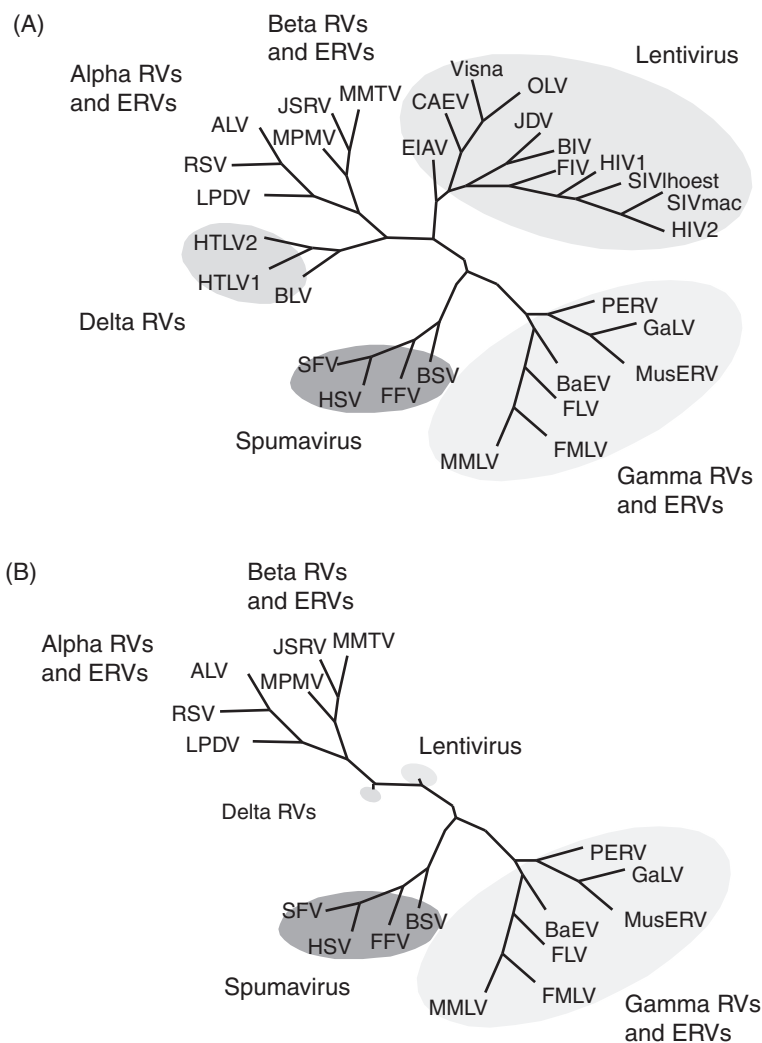


FIGURE 13.4 (A) Standard Protopars tree for the retroviral integrase sequence. Shading helps distinguish the retroviral families. (B) The lentiviral and HTLV lineages only have been shortened to take into account their much faster fixation rates compared to that for the spumavirus lineages. It is tempting to conclude that in a temporal sense the lentiviruses and delta retroviruses have recently emerged from some sort of endogenous retroviral baseline. (See Plate 16 for the color version of this figure.)

However, there is a complication. Between retroviruses, fixation rates vary hugely. For the lentiviruses, read HIV/SIV simply because the databases are so much larger, the amino acid fixation rates are of the order of 0.1–1% per year depending on the protein (Johnson *et al.*, 1991). For HTLV overall fixation rates are more like 1% per century while for the foamy viruses or spumaviruses it is estimated that the rates could be as slow as ~2% per million years (Switzer *et al.*, 2005). Obviously this cannot be ascribed to the RTase as the HTLV point mutation rate is $\sim 7 \times 10^{-6}$ per base per cycle (Mansky, 2000) which is only four-fold down on that for HIV ($\sim 3 \times 10^{-5}$). It transpires that viruses of the HTLV family replicate a little by reverse transcription and extensively by Tax-driven clonal expansion of the host cell, i.e. mitosis. Tax is a viral protein that impacts the cell by tripping it into cycling even during the non-malignant carrier state. Quite why the foamy viruses fix mutations so slowly has not been clearly worked out, yet it might well involve clonal expansion, as the consensus opinion is that retrovirus and RNA polymerase complexes are not accompanied by proofreading enzymes.

Nonetheless the consequences for phylogeny are considerable. Understandably a phylogenetic tree being built essentially from a distance matrix, is a scalar construction. Yet in our mind's eye it is inevitably read in a temporal sense, which is also understandable, for after all we are interested in evolution, i.e. dx/dt . There is the implicit assumption that lineages are lengthening at roughly equal rates. However, when the rates differ by several orders of magnitude scalar trees hide information. If the lentiviral and HTLV and normalized to the spumavirus lineage as a function of their fixation rates, then both the large and diverse lentivirus group as well as the small HTLV group collapse to next to nothing (Figure 13.4B). In this tree, the other retroviral lineages have been kept constant either for lack of data about fixation rates, or else the possibility of recombination with endogenous retroviruses is such as to confound any evaluation.

How are we to interpret the hobbled lentivirus and HTLV viruses? As mere blips in retrovirology? If so they certainly represent a recent blip from some sort of a retroviral base line. Yet given the complexity of lentiviral genomes, considerable changes in synteny have occurred in a very short time compared to the relatively fewer differences in the spumavirus group over much greater periods of time.

HOW FAR CAN VARIATION GO FOR A RETROVIRUS?

Given the phenomenal fixation rates, particularly for the lentiviruses, the question often asked is, are there limits to the change a virus can absorb? This was dealt with extensively in my chapter of the original edition of this book and so there is no need to belabor the point—functional sequence space is so vast that there are effectively no limits. One example will suffice: feline immunodeficiency virus produces profound immunodeficiency and an AIDS-like disease in domestic and big wild cats. A simple alignment of the envelope polypeptide precursor from the PPR and Pallas cat isolates shows a stunning 10% match for the complete Env sequence (Figure 13.5) and approaches the 5% expectation value given 20 amino acids. Despite this both bear all the hallmarks of a lentiviral envelope sequence—they are highly glycosylated, rich in cysteine residues, have a long cytoplasmic tail and contain the requisite three hydrophobic segments. In general for retroviral proteins, the envelope proteins are the most divergent while the enzymes encoded by *pol* are the least variable. Thus the rate of fixation of integrase is perhaps a factor of 10 less than for envelope (Johnson *et al.*, 1991). So long as the protein fold and a few crucial residues are preserved, the rest will, and does, change given time.

To appreciate this it is worth mentioning a contrary situation. The “Paracelsus challenge” defied or encouraged researchers to radically change the fold of a known protein by changing less than 50% of its amino acid

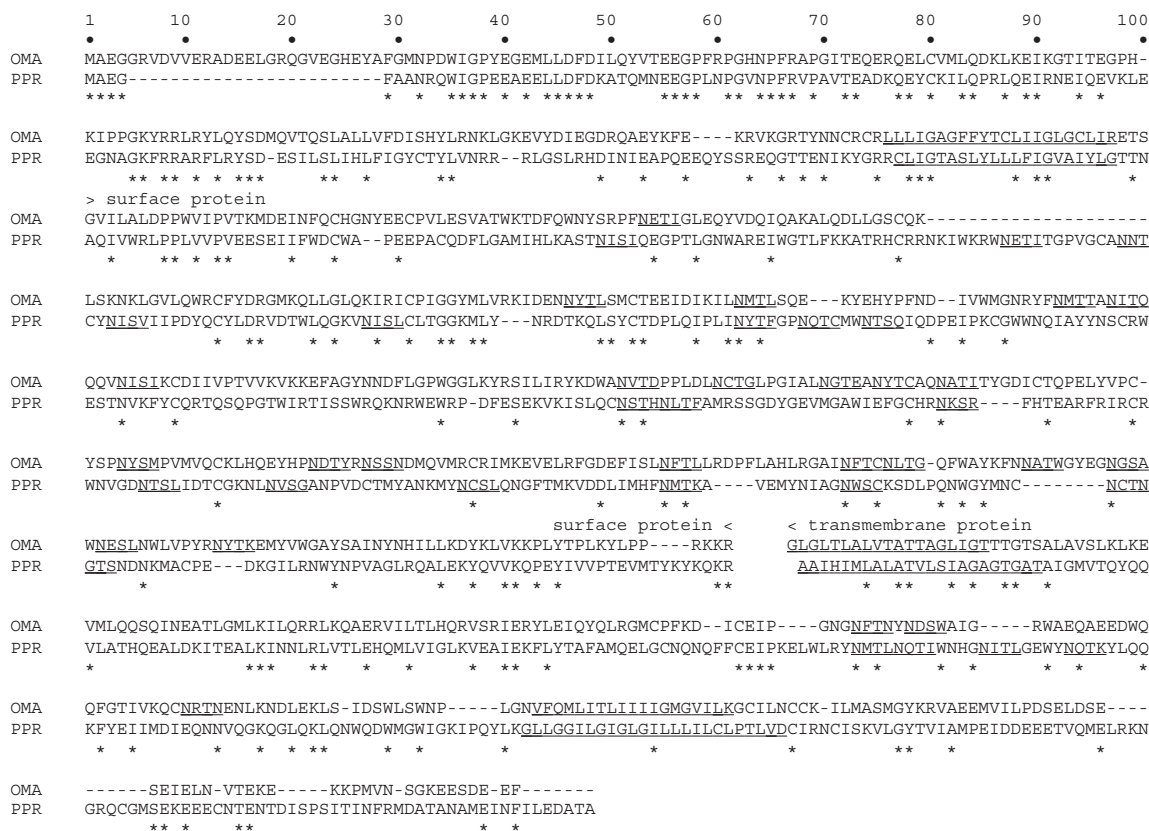


FIGURE 13.5 Envelope variation for two strains of feline immunodeficiency virus (FIV). The accession numbers for the PPR and Pallas cat OMA strains are M36968 and U56928 respectively. Amino acid identities are scored as asterisks. The membrane spanning segments are underlined, as are the *N*-linked glycosylation sites. Despite the differences, which include extensive changes in the positions of the disulfide bonds, both were derived from viable strains.

residues (Rose and Creamer, 1994). The challenge was first met in 1997. The $\beta 1$ domain of the *Streptococcal* IgG-binding protein G which displays an extensive β -sheet structure was redesigned and transformed into the *E. coli* protein Rop, a homodimeric four-helix bundle protein (Dalal *et al.*, 1997). The first solution was perhaps not surprisingly called Janus. This designer protein is a fascinating example of cultural evolution impinging on biology. It represents a statistically improbable saltation as opposed to, the familiar and presumably “incremental” polymerase mutations that allow FIV Env to drift beyond way beyond the 50% level separating Janus from its natural counterpart.

PHYLOGENY

Retrovirology has been a hotbed of methodological improvement in the realm of phylogeny over the past decade. A rich variety of methods have emerged; those which have been particularly useful being those enabling the detection of complex recombinant structures. Despite this problems remain. Many simple assumptions and corrections haven’t been developed using real data sets. For example, a complete retrovirus tree includes a branch for the lentiviruses and the δ -retroviruses (HTLV-BLV group) that have profound differences in base composition, codon usage, mutation matrices, and fixation rates as

mentioned above. Applying corrections for back mutations is not easy because there are hardly any good data sets on which to train methods.

Recombination is a major problem for sequences jump repeatedly, rather than gradually diffuse or radiate through sequence space. Of course sequence space is of such high dimensionality that words essentially fail us, words like jump, diffuse, and radiate being used vainly to communicate some sort of sense. Recombination introduces homoplasies that are otherwise relatively rare (Pelletier *et al.*, 1995; Wain-Hobson *et al.*, 2003). Ignoring these tends to inflate the number of mutations in a data set, which means that the sequences would appear older than they really were. Unfortunately network analyses that allow a clear description of recombination work only on small numbers of informative sites and small data sets. Nonetheless they show that recombination can sometimes account for up to 50% of all substitutions in a sequence set (Wain-Hobson *et al.*, 2003). For HIV, the simplest conclusion was that the number of true mutations in a set of PCR-derived sequences is equal to the number of variable sites.

An additional problem plagues HIV phylogeny. It is clear that relatively old proviruses can be reactivated from the carrying T cell, presumably by antigenic stimulation (Bello *et al.*, 2005). This means that not all progeny are derived from the most recently infected cells. This would serve to reduce the intra-patient cross-sectional diversity, and perhaps underestimate the time of infection. When sequences are close they are scored as such by whatever phylogenetic method. Equally, when sequences are highly divergent they remain so. However, interpreting deep relationships or asking too precise a question is going to be fraught with assumptions.

GENOME DESIGN IMPACTS THE TEMPO OF EVOLUTION

The design of a retroviral genome can strongly impact replication and hence evolution. It has

been known for some years that the HTLV-1 protein Tax transactivates a number of host cell genes, among them the IL-2 and IL-2 receptor genes helping set up an autocrine system (Yoshida, 1993). It has been shown that Tax intervenes directly in cell proliferation by inhibiting the cyclin D/Rb/p16^{INK4A} pathway. A consequence of this, clonal expansion of HTLV-1 provirus-bearing T cells, has recently been shown via PCR-based studies of the HTLV-1 integration sites (Wattel *et al.*, 1995). In asymptomatic carriers and patients with neurological disease devoid of malignancy, clonal expansion was found to be the norm. That HTLV-1 replicates mainly via mitosis reconciles the problem of the simultaneous occurrence of high proviral load with genetic stability as proviral synthesis is accomplished essentially by the host cell replication machinery, which is endowed with proofreading systems.

However, this raises another problem: how does a clonally expanded infected cell produce little virus? Here the answers are less complete. Intrinsically, HTLV-1 production is low compared with that of HIV and is an inevitable consequence of their different genome organization (Figure 13.6). Both HIV and HTLV-1 encode analogous viral transcription transactivators, Tat and Tax, and temporal regulators of splicing, Rev and Rex respectively. The regulator proteins have a negative effect upon Tat and Tax which also curtails expression of the genomic RNA, and the large mRNA encoding the key virion proteins, Gag, Pol, and Env. HIV encodes both a Rev-dependent and -independent *tat* mRNA (Figure 13.6) allowing continual Tat-driven transcription and Rev co-ordinated splicing, and consequently continued expression of virion, *gag* and *pol* mRNAs. By contrast, the HTLV-1 analogues Tax and Rex are encoded by the same mRNA (Figure 13.6). Thus Rex tends to increase Gag, Pol, and Env expression at the expense of both Tax and Rex. The result is an equilibrium with greatly reduced virus production. In short, via genome organization, molecular switches with negative feedback loops can be generated which grossly

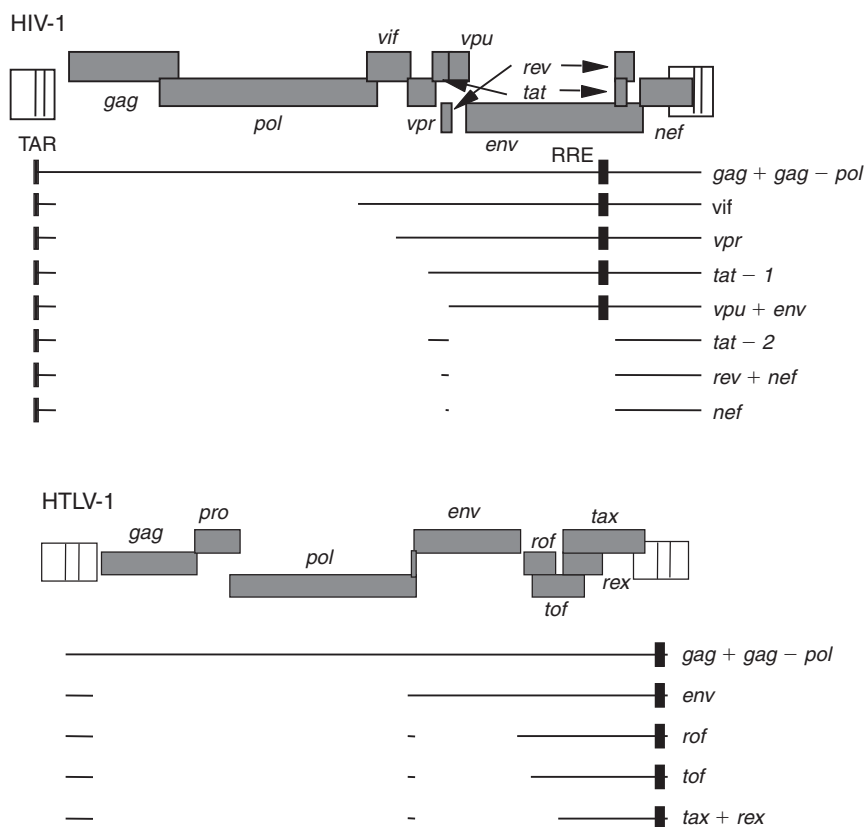


FIGURE 13.6 Transcription maps of HTLV-1 and HIV-1. Below the genome configuration is the transcription map, each line designating a unique mRNA. When broken, the mRNA is made up of ≥ 2 exons. The HTLV-1 transactivator, Tax, acts on a DNA element in LTR U3 region and cellular genes. That of HIV, Tat, recognizes the RNA TAR element in the R region of the LTR. The analogous HTLV-1 and HIV-1 regulators of splicing, Rex and Rev, act on specific intragenomic RNA structures, RxRE and RRE respectively. Note that Tax and Rex are derived from the same mRNA. The HIV-1 Tat protein comes in two functionally equivalent forms of 72 and 86 residues. The 86 residue form (exons 2+3) comes from a small Rev-independent mRNA while the 76 residue form (exon 2) comes from a large Rev-dependent mRNA.

impact the way a virus replicates and evolves. One final point must be added; the Tax-transactivated HTLV-1 promoter is much less powerful than the Tat-transactivated HIV-1 LTR.

ENDOGENOUS RETROVIRUSES AND MULTIPLE INTRODUCTIONS

Within the phylogenetic analysis above (Figure 13.4) are featured a number of complete endogenous retrovirus (ERV) sequences. These range

from recently inactivated proviruses such as the HERV-Ks to baboon endogenous retrovirus (BaEV) that circulates in the peripheral blood of baboons. The retrovirus is produced but unable to infect its own cells, a simple example of negative selection against insertional gene inactivation. As can be seen the majority of these mammalian ERVs are clustered into two groups for which there are exogenous retroviral counterparts. Massive PCR-based efforts have shown that the extent of ERVs spreads across a vast spectrum of the living world, from insects, turtles, snakes, and of course

mammals (Gifford and Tristem, 2003; Belshaw *et al.*, 2004, 2007; Gifford *et al.*, 2005). There is no doubt that retroviruses have invaded germinal cells many times in the past. Equally there is no doubt that some ERVs can escape their host and infect another species, the case of gibbon ape leukemia virus (GaLV) being exemplary. Its closest relative is a murine ERV (Wolgast *et al.*, 1998). While there were numerous reports of endogenous counterparts to the lentiviruses, foamy virus, the alignments were borderline (Cordonnier *et al.*, 1995).

A recent and glorious surprise is afforded by the aptly named sequence RELIK, for rabbit endogenous lentivirus, where the K refers to the tRNA lysine (K) primer used to initiate reverse transcription (Katzourakis *et al.*, 2007). Through a bioinformatics analysis a number of extensively degraded copies of an endogenous retrovirus were identified in the rabbit genome. After careful reconstruction of a consensus sequence it transpired that these relics were those of an endogenous lentivirus, the first unambiguous example to date (Figure 13.1). A simple comparison of the genome organization with named orfs is enough to convince. What is so nice about RELIK are the fine details that any lentivirologist would pick up in a moment. For example, the envelope protein is highly glycosylated, full of cysteine residues, has a furin-like cleavage site and encodes a long intracytoplasmic tail. It encodes a dUTP sequence precisely located between RTase/RNaseH and integrase (Figure 13.2). The primer-binding site (tRNA_{Lys3}) and polypurine tract are typical of a lentivirus.

Despite all the above traits, in terms of gene organization it is the least complex lentivirus genome. Apart from *gag*, *pol*, and *env*, there are only the *tat* and *rev* genes. There are none of the so-called, yet misnamed, accessory genes typical of the primate lentiviruses. It is tempting to consider RELIK as a precursor to the lentiviruses, although one must be ever careful of the possibility of loss of function. A phylogenetic analysis based on the RTase/integrase domains shows it to be the closest relative to the lentiviruses, yet it is the outlier to all extant lentiviruses. Given that RELIK is the

outlier, it is plausible that the monophyletic lentiviral group is derived from some ancestor of this lentivirus. This concords with the relative simplicity of the genetic organization of RELIK.

From a detailed analysis of the integration sites, the authors estimate that the invasion of the rabbit germline might have taken place some 7 million years ago. As there is no reason to doubt this finding we are placed in a very interesting evolutionary spot. If the present set of lentiviruses have remained exogenous for so long, then from what is known of the fixation rates of the primate lentiviruses, of the order of 0.1% amino acids replacements per year for the most conserved integrase gene, then it should not be possible to find 40% amino acid sequence identity between the HIV-1 and RELIK integrases. In the light of RELIK it seems more parsimonious to assume that the lentiviruses are very recently derived from an endogenous lentiviral RV and have been mutating furiously ever since. As mammalian genomics studies develop rapidly, the hope is that a much less degraded endogenous retrovirus will turn up in the near future.

BACK TO THE FUTURE—OR RECOVERING RETROVIRAL ANCESTORS

Even though the RELIK sequences were highly degraded, a consensus sequence was derived that is qualitatively satisfying. The ultimate would be to synthesize the provirus in the hope of resurrecting a functional lentivirus. Successful recovery of an active human ERV (HERV-K) was recently achieved following mutagenesis of a minimally defective HERV-K molecular clone (Lee and Bieniasz, 2007). Other workers have been calculating consensus or ancestral sequences for the HIV-1 envelope protein in the hope of making an immunogen that is equidistant to a maximum number of strains circulating worldwide. Chemical synthesis of the entire

coding sequence and cloning into an expression vector generated functional envelope proteins that were remarkably "HIV-like" (Nickle *et al.*, 2003; Rolland *et al.*, 2007; Kothe *et al.*, 2007). Of course, the ultimate in this area is the complete chemical syntheses of an infectious molecular clone of poliovirus (Cello *et al.*, 2002) and subsequently Φ X174.

The interesting evolutionary point here, which is a general one, is that total chemical synthesis of a virus, or "therapy," of a partly degraded retrovirus constitutes a hiatus in Darwinian evolution. The mantra has always been "descent with modification," and with reason. Otherwise stated there is an uninterrupted series of DNA replication rounds linking this author to one of the earliest single-celled DNA organisms floating around 3 billion years ago. Total chemical synthesis takes biological evolution into the realm of cultural evolution, something totally new. While the resulting viruses *are* poliovirus and Φ X174, from an evolutionary setting where biology and history are inseparable, they are worlds apart. No other discipline is as advanced. The bacteriologists threaten to make a "minimalist bacterium," but so far they are not there.

In fact ever since 1962 and the demonstration that phenol-extracted papillomavirus DNA could be infectious (Ito, 1962; Chambers and Ito, 1964), virologists have been playing games with the Darwinian mantra.

HOW GOOD IS THE RETROVIRAL TREE?

Given that retroviral replication is endlessly interrupted by recombination, while many of the corrections applied to phylogeny are rather ad hoc and approximate, is it possible that the phylogeny is completely wrong? How can it be checked? It is possible to encode retroviral organization in terms of a series of answers to questions that command discrete responses, for example yes or no, that can be coded as 1 or 0. Does *env* overlap *pol*? Does the genome encode a *bet* orf? Features

like the density of glycosylation sites in the Env protein, which in fact varies hugely, represent a continuous variable which is not useful. Questions with discrete answers have the advantage of being independent of the precise nucleotide sequence, polymorphisms, multiple substitutions, and saturation.

When this is done for the retroviruses (Renoux-Elbe *et al.*, 2002) it turns out that the arborescence is reasonably similar to that provided by the retroviral integrase gene, although there are a few differences (Figure 13.7). However, overall it is a comforting result. The answers can be overlaid on the non-classical tree (Figure 13.8). The minimal retroviral organization was always *gag-pol-env* and so there are no surprises that this configuration exemplified by MLV is found branching near the base of the tree. However, the lentiviruses turn out to be the largest and most varied tree. Indeed there is almost as much novelty within this group as in all the others and throws into light the trivial distinctions that were used in the past to lump or split retroviruses. The minimal lentivirus turns out to be *gag-pol-tat-rev-env* with the acquisition of *vif* coming immediately after. However, as all but one lentivirus (EIAV) encodes a *vif* gene, this could be a legitimate case of loss of function. Since the tree in Figure 13.8 was calculated, "RELIX" was discovered (Katzourakis *et al.*, 2007). As mentioned above, RELIX has this minimalist lentivirus organization, i.e. *gag-pol-tat-rev-env* yet does not encode a *vif* gene. This is consistent with the hypothesis that *vif* was a subsequent acquisition compared to *tat* and *rev*.

As always care has to be exercised when interpreting trees. Basically sequence space is so huge and multidimensional that simple two-dimensional constructs cannot do justice to the complexity. While not shown here, many of the bootstrap values are rather small, indicating that parts of the tree, particularly towards the root, are not robust (Renoux-Elbe *et al.*, 2002). It should perhaps be mentioned that the endogenous retrovirus HERV-K encodes a *rev*-like orf. It is of course tempting to conclude that *rev* appeared before *tat*.

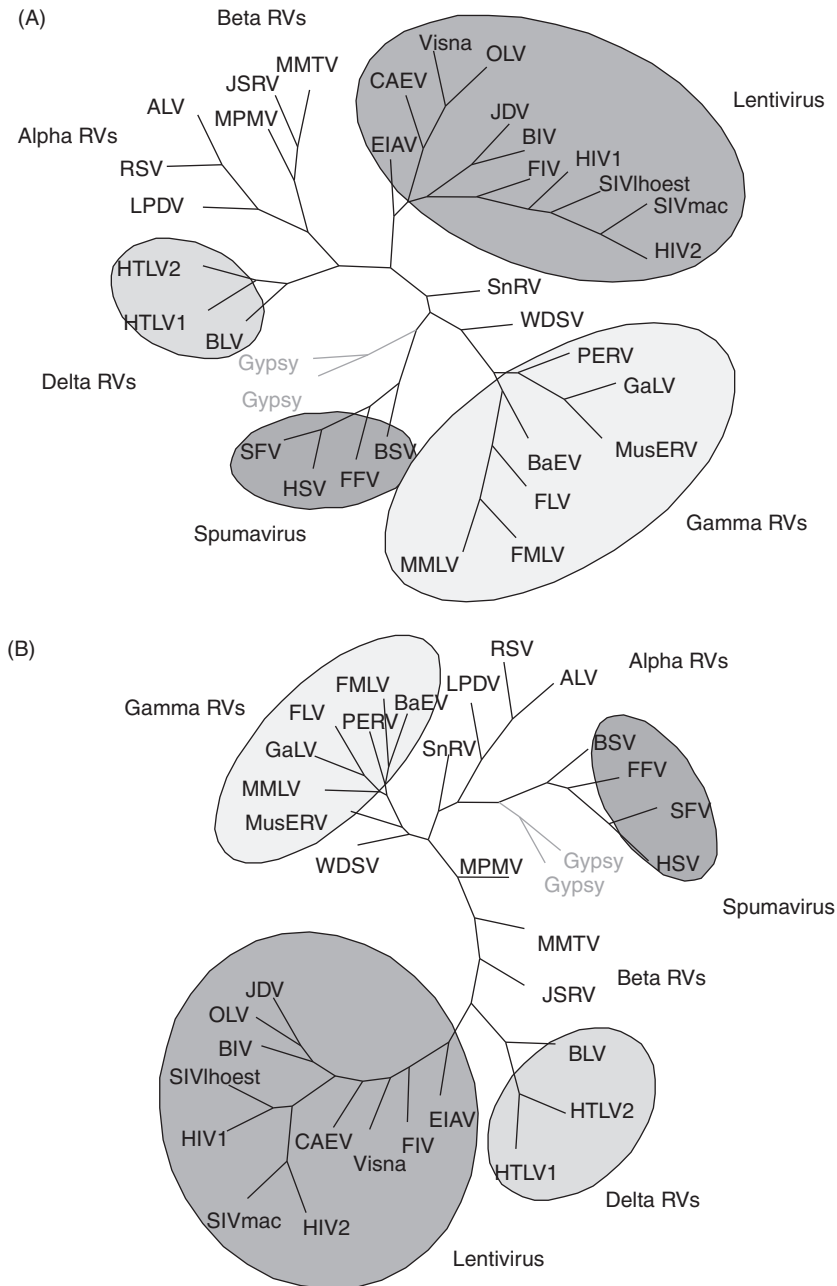


FIGURE 13.7 Classical and synteny-based retroviral trees. (A) Classical Protopars tree for the retroviral integrase sequence. The shaded ellipses highlight the different retroviral groups. Two *Gypsy* retroviral elements from *Drosophila* are used as outgroups. (B) A syntenic tree based on a series of questions giving binary (1, 0) answers. While there is general agreement, the positions of the alpha- and betaretroviruses vary. (See Plate 17 for the color version of this figure.)

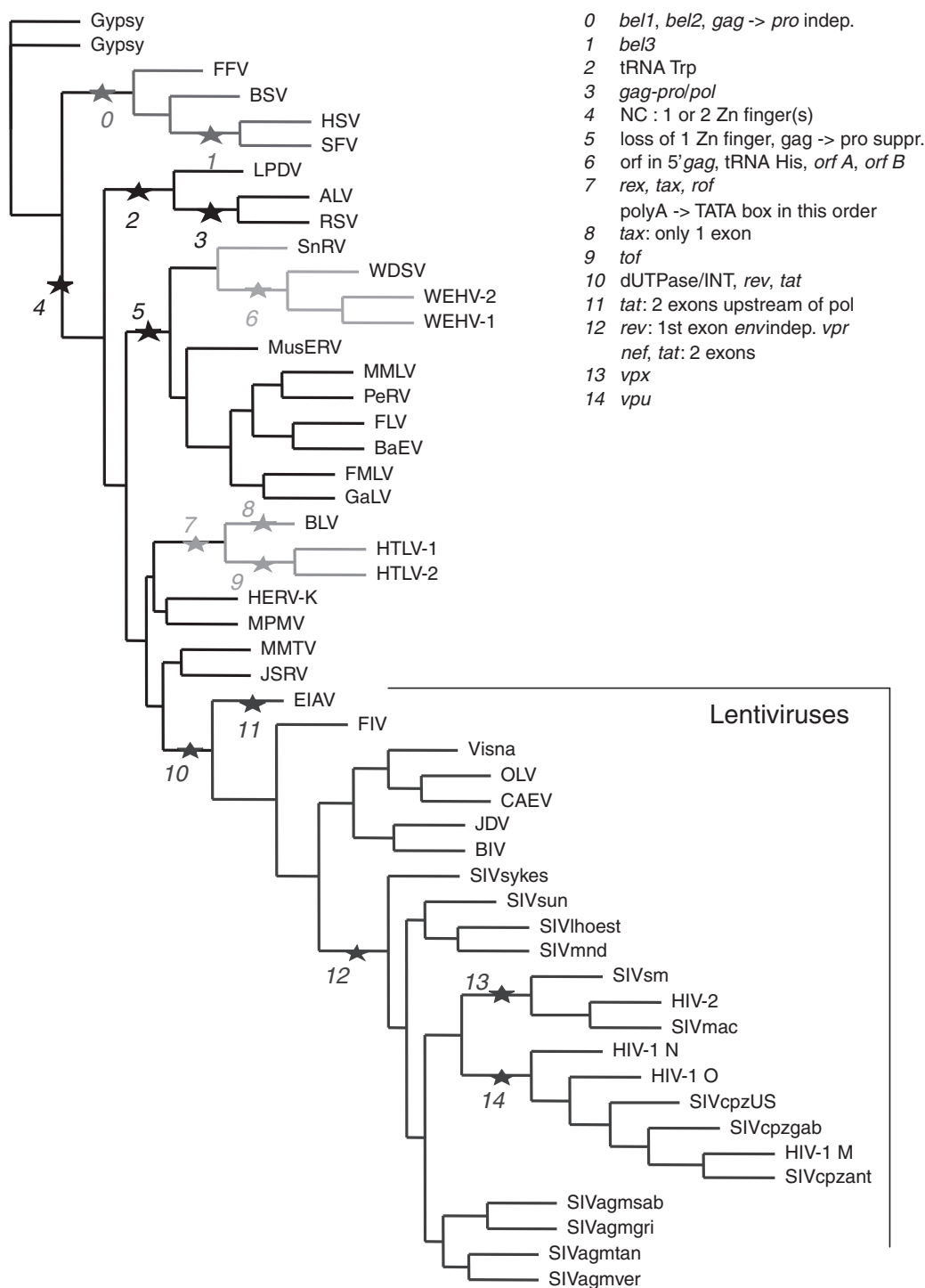


FIGURE 13.8 Annotated phylogenetic retroviral tree based on the binary answer mode alignment. Each of the 15 stars corresponds to one or more characters, i.e. organizational traits or evolutionary events at the origin of one group of retroviruses. The lentiviral group represents the largest group of retroviruses in terms of genome complexity. (See Plate 18 for the color version of this figure.)

While it is satisfying that HERV-K is found in the twilight zone not too far from the lentiviruses, given the low bootstrap values it is not possible to be more precise.

THE FUTURE OF RETROVIRAL EVOLUTION?

An additional point can be made concerning Figures 13.7 and 13.8. The standard retroviral tree concerns the accumulation of point mutations while the “unconventional” tree concerns the accumulation of novel features. That the general branching orders coincide suggests that the acquisition of novelty *inevitably* accompanies the remorseless accumulation of point mutations. As any retrovirus radiates through sequence space it will inevitably accumulate novel features along with massive numbers of point mutations. As the lentiviruses sport the highest fixation rates of any retrovirus, organizational novelty is to be expected here, and probably with the accompanying HIV-1 pandemic more than anywhere else.

It transpires that almost every small equatorial African monkey harbors a distinct SIV, albeit pathogenic. To date there are >34 SIVs if not more (Van Heuverswyn *et al.*, 2006). Through hunting and butchering it is eminently reasonable to postulate yet another zoonotic transmission, and perhaps secondary transmission between humans. While a HIV-3 cannot be ruled out, it seems hard to beat HIV-1 that has in the space of a half-century climbed to the top of the virological Richter scale (Hale *et al.*, 2001; Weiss and McMichael, 2004).

If one of these small monkey SIVs ever crosses to humans another pandemic is not axiomatic. There is one unfortunate “experiment” that was unwittingly performed on humans and yet, fortunately, did not take off. With the identification of SV40 in rhesus macaque primary kidney cultures, manufacture of the attenuated Sabin strains of poliovirus was switched to African green monkey (Agm) kidney cultures. It is no secret that the incidence of SIVagm in some troops of Agms

can be as high as 70%. As there was no inactivating step in the vaccine preparation, some individuals must have received SIVagm. Yet HIV-1 simply is not derived from any SIVagm of any sort despite the fact that some strains can grow on human peripheral blood mononuclear cells (Gautam *et al.*, 2007). It was, as Robin Weiss remarked, a close shave.

The only real question of any importance for us as potential hosts for retrovirus replication is whether variation will be the Achilles’ heel of an eventual HIV vaccine. Is it going to be a problem? Certainly studies show that an attenuated SIV vaccine can contain the challenge virus despite a certain degree of envelope variation (Johnson *et al.*, 1999; Blancou *et al.*, 2004). However, breakthrough does occur when more diverse challenge strains are used. There are simply not enough data to be more precise on the degree of variation. Superinfection by HIV-1 is often given as proof that vaccination will not work. This is not the same as vaccination. It is increasingly clear that during the very early stages of infection there is a loss of HIV/SIV-specific immunity, meaning that a seropositive individual is not the same as a vaccinated volunteer. Whatever the state of experimental findings and subtending riders, HIV variation is not good news.

To conclude, retroviruses are remarkable mutation machines and their genetic complexity and lifestyles vary enormously. Almost anything is possible given time. With the massive effort devoted to sequencing complete genomes, more endogenous retroviruses will be described in less familiar species, and why not clear counterparts to HTLV and spumaviruses? The koala endogenous retrovirus is an interesting case in point (Gifford, 2006), although there must be many more. Although there are a handful of retroviruses in fish, the aquatic world is so diverse, that there must be many more. Concerning extant retroviruses, the lentiviruses represent the most diverse or complex group. As their fixation rates are so high, the probability is higher that new surprises will show up within this group rather than others. The biologist J.D. Bernal remarked almost 50 years ago that “everything that can

happen will happen, and nobody will be safe from it." While I would not share the rather dark 3' end to his remark because we now can, after all, vaccinate against a large number of microbes and successfully treat one human cancer out of two, the 5' part is a good précis of retrovirus evolution.

REFERENCES

- Ablashi, D.V., Eastman, H.B., Owen, C.B., Roman, M.M., Friedman, J., Zabriskie, J.B. *et al.* (2000) Frequent HHV-6 reactivation in multiple sclerosis (MS) and chronic fatigue syndrome (CFS) patients. *J. Clin. Virol.* **16**, 179–191.
- Bello, G., Casado, C., Sandonis, V., Alonso-Nieto, M., Vicario, J.L., Garcia, S. *et al.* (2005) A subset of human immunodeficiency virus type 1 long-term non-progressors is characterized by the unique presence of ancestral sequences in the viral population. *J. Gen. Virol.* **86**, 355–364.
- Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Paces, J., Burt, A. and Tristem, M. (2004) Long-term reinfection of the human genome by endogenous retroviruses. *Proc. Natl Acad. Sci. USA* **101**, 4894–4899.
- Belshaw, R., Watson, J., Katzourakis, A., Howe, A., Woolven-Allen, J., Burt, A. and Tristem, M. (2007) Rate of recombinational deletion among human endogenous retroviruses. *J. Virol.* **81**, 9437–9442.
- Bibollet-Ruche, F., Gao, F., Bailes, E., Saragosti, S., Delaporte, E., Peeters, M. *et al.* (2004) Complete genome analysis of one of the earliest SIVcpzPtt strains from Gabon (SIVcpzGAB2). *AIDS Res Hum Retroviruses* **20**, 1377–1381.
- Blancou, P., Chenciner, N., Ho Tsong Fang, R., Monceaux, V., Cumont, M.C., Guetard, D. *et al.* (2004) Simian immunodeficiency virus promoter exchange results in a highly attenuated strain that protects against uncloned challenge virus. *J. Virol.* **78**, 1080–1092.
- Cello, J., Paul, A.V. and Wimmer, E. (2002) Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* **297**, 1016–1018.
- Chambers, V.C. and Ito, Y. (1964) Morphology of Shope papilloma virus associated with nucleic acid-induced tumors of cottontail rabbits. *Virology* **23**, 434–436.
- Coffin, J.M. (1979) Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses. *J. Gen. Virol.* **42**, 1–26.
- Cordonnier, A., Casella, J.F. and Heidmann, T. (1995) Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *J. Virol.* **69**, 5890–5897.
- Dalal, S., Balasubramanian, S. and Regan, L. (1997) Protein alchemy: changing beta-sheet into alpha-helix. *Nat. Struct. Biol.* **4**, 548–552.
- Delaroque, N., Boland, W., Muller, D.G. and Knippers, R. (2003) Comparisons of two large phaeoviral genomes and evolutionary implications. *J. Mol. Evol.* **57**, 613–622.
- Delebecque, F., Suspene, R., Calattini, S., Casartelli, N., Saib, A., Froment, A. *et al.* (2006) Restriction of foamy viruses by APOBEC cytidine deaminases. *J. Virol.* **80**, 605–614.
- Drake, J.W. and Holland, J.J. (1999) Mutation rates among RNA viruses. *Proc. Natl Acad. Sci. USA* **96**, 13910–13913.
- Gautam, R., Carter, A.C., Katz, N., Butler, I.F., Barnes, M., Hasegawa, A. *et al.* (2007) In vitro characterization of primary SIVsmm isolates belonging to different lineages. In vitro growth on rhesus macaque cells is not predictive for in vivo replication in rhesus macaques. *Virology* **362**, 257–270.
- Gifford, R.J. (2006) Evolution at the host-retrovirus interface. *Bioessays* **28**, 1153–1156.
- Gifford, R. and Tristem, M. (2003) The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* **26**, 291–315.
- Gifford, R., Kabat, P., Martin, J., Lynch, C. and Tristem, M. (2005) Evolution and distribution of class II-related endogenous retroviruses. *J. Virol.* **79**, 6478–6486.
- Hale, P., Makgoba, M.W., Merson, M.H., Quinn, T.C., Richman, D.D., Vella, S. *et al.* (2001) Mission now possible for AIDS fund. *Nature* **412**, 271–272.
- Harris, R.S. and Liddament, M.T. (2004) Retroviral restriction by APOBEC proteins. *Nat. Rev. Immunol.* **4**, 868–877.
- Ito, Y. (1962) Relationship of components of papilloma virus to papilloma and carcinoma cells. *Cold Spring Harb. Symp. Quant. Biol.* **27**, 387–394.
- Jetzt, A.E., Yu, H., Klarmann, G.J., Ron, Y., Preston, B.D. and Dougherty, J.P. (2000) High rate of recombination throughout the human immunodeficiency virus type 1 genome. *J. Virol.* **74**, 1234–1240.
- Jin, M.J., Hui, H., Robertson, D.L., Muller, M.C., Barre-Sinoussi, F., Hirsch, V.M. *et al.* (1994) Mosaic genome structure of simian immunodeficiency virus from west African green monkeys. *EMBO J.* **13**, 2935–2947.
- Johnson, P.R., Hamm, T.E., Goldstein, S., Kitov, S. and Hirsch, V.M. (1991) The genetic fate of molecularly cloned simian immunodeficiency virus in experimentally infected macaques. *Virology* **185**, 217–228.
- Johnson, R.P., Lifson, J.D., Czajak, S.C., Cole, K.S., Manson, K.H., Glickman, R. *et al.* (1999) Highly attenuated vaccine strains of simian immunodeficiency virus protect against vaginal challenge: inverse relationship of degree of protection with level of attenuation. *J. Virol.* **73**, 4952–4961.
- Jung, A., Maier, R., Vartanian, J.P., Bocharov, G., Jung, V., Fischer, U. *et al.* (2002) Multiply infected spleen cells in HIV patients. *Nature* **418**, 144.
- Katzourakis, A., Tristem, M., Pybus, O.G. and Gifford, R.J. (2007) Discovery and analysis of the first endogenous lentivirus. *Proc. Natl Acad. Sci. USA* **104**, 6261–6265.

- Kawai, S. and Hanafusa, H. (1972) Genetic recombination with avian tumor virus. *Virology* **49**, 37–44.
- Kothe, D.L., Decker, J.M., Li, Y., Weng, Z., Bibollet-Ruche, F., Zammit, K.P., Salazar, M.G. *et al.* (2007) Antigenicity and immunogenicity of HIV-1 consensus subtype B envelope glycoproteins. *Virology* **360**, 218–234.
- Lee, Y.N. and Bieniasz, P.D. (2007) Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog.* **3**, e10.
- Mahieux, R., Suspene, R., Delebecque, F., Henry, M., Schwartz, O., Wain-Hobson, S. and Vartanian, J.P. (2005) Extensive editing of a small fraction of human T-cell leukemia virus type 1 genomes by four APOBEC3 cytidine deaminases. *J. Gen. Virol.* **86**, 2489–2494.
- Mansky, L.M. (2000) In vivo analysis of human T-cell leukemia virus type 1 reverse transcription accuracy. *J. Virol.* **74**, 9525–9531.
- Mansky, L.M. and Temin, H.M. (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.* **69**, 5087–5094.
- Mansky, L.M. and Wisniewski, R.M. (1998) The bovine leukemia virus encapsidation signal is composed of RNA secondary structures. *J. Virol.* **72**, 3196–3204.
- McGeoch, D.J. (1990) Protein sequence comparisons show that the ‘pseudoproteases’ encoded by poxviruses and certain retroviruses belong to the deoxyuridine triphosphatase family. *Nucleic Acids Res.* **18**, 4105–4110.
- Mikl, M.C., Watt, I.N., Lu, M., Reik, W., Davies, S.L., Neuberger, M.S. and Rada, C. (2005) Mice deficient in APOBEC2 and APOBEC3. *Mol. Cell. Biol.* **25**, 7270–7277.
- Nickle, D.C., Jensen, M.A., Gottlieb, G.S., Shriner, D., Learn, G.H., Rodrigo, A.G. and Mullins, J.I. (2003) Consensus and ancestral state HIV vaccines. *Science* **299**, 1515–1518.
- Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C. *et al.* (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**, 171–182.
- Peeters, M., Liegeois, F., Torimiro, N., Bourgeois, A., Mpoudi, E., Vergne, L. *et al.* (1999) Characterization of a highly replicative intergroup M/O human immunodeficiency virus type 1 recombinant isolated from a Cameroonian patient. *J. Virol.* **73**, 7368–7375.
- Pelletier, E., Saurin, W., Cheynier, R., Letvin, N.L. and Wain-Hobson, S. (1995) The tempo and mode of SIV quasispecies development in vivo calls for massive viral replication and clearance. *Virology* **208**, 644–652.
- Renoux-Elbe, C., Cheynier, R. and Wain-Hobson, S. (2002) Phylogeny derived from coding retroviral genome organization. *J. Mol. Evol.* **54**, 376–385.
- Rolland, M., Jensen, M.A., Nickle, D.C., Yan, J., Learn, G.H., Heath, L. *et al.* (2007) Reconstruction and function of ancestral center-of-tree human immunodeficiency virus type 1 proteins. *J. Virol.* **81**, 8507–8514.
- Rose, G.D. and Creamer, T.P. (1994) Protein folding: predicting. *Proteins* **19**, 1–3.
- Sheehy, A.M., Gaddis, N.C., Choi, J.D. and Malim, M.H. (2002) Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418**, 646–650.
- Stehelin, D., Varmus, H.E., Bishop, J.M. and Vogt, P.K. (1976) DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170–173.
- Suspene, R., Guetard, D., Henry, M., Sommer, P., Wain-Hobson, S. and Vartanian, J.P. (2005) Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo. *Proc. Natl Acad. Sci. USA* **102**, 8321–8326.
- Switzer, W.M., Salemi, M., Shanmugam, V., Gao, F., Cong, M.E., Kuiken, C. *et al.* (2005) Ancient co-speciation of simian foamy viruses and primates. *Nature* **434**, 376–380.
- Tristem, M., Marshall, C., Karpas, A., Petrik, J. and Hill, F. (1990) Origin of vpx in lentiviruses. *Nature* **347**, 341–342.
- Van Heuverswyn, F., Li, Y., Neel, C., Bailes, E., Keele, B.F., Liu, W. *et al.* (2006) Human immunodeficiency viruses: SIV infection in wild gorillas. *Nature* **444**, 164.
- Vandvik, B. and Norrby, E. (1989) Paramyxovirus SV5 and multiple sclerosis. *Nature* **338**, 769–771.
- Vogt, P.K. (1971) Genetically stable reassortment of markers during mixed infection with avian tumor viruses. *Virology* **46**, 947–952.
- Wain-Hobson, S., Renoux-Elbe, C., Vartanian, J.P. and Meyerhans, A. (2003) Network analysis of human and simian immunodeficiency virus sequence sets reveals massive recombination resulting in shorter pathways. *J. Gen. Virol.* **84**, 885–895.
- Wattel, E., Vartanian, J.P., Pannetier, C. and Wain-Hobson, S. (1995) Clonal expansion of human T-cell leukemia virus type I-infected cells in asymptomatic and symptomatic carriers without malignancy. *J. Virol.* **69**, 2863–2868.
- Weir, J.P. (1998) Genomic organization and evolution of the human herpesviruses. *Virus Genes* **16**, 85–93.
- Weiss, R.A. and McMichael, A.J. (2004) Social and environmental risk factors in the emergence of infectious diseases. *Nat. Med.* **10**, S70–76.
- Weiss, R.A., Mason, W.S. and Vogt, P.K. (1973) Genetic recombinants and heterozygotes derived from endogenous and exogenous avian RNA tumor viruses. *Virology* **52**, 535–552.
- Winkelmann, J.C. (1993) HTLV-1 and multiple sclerosis: the link is missing. *J. Lab. Clin. Med.* **122**, 230–231.
- Wolgamot, G., Bonham, L. and Miller, A.D. (1998) Sequence analysis of *Mus dunni* endogenous virus reveals a hybrid VL30/gibbon ape leukemia virus-like structure and a distinct envelope. *J. Virol.* **72**, 7459–7466.
- Yoshida, M. (1993) HTLV-1 Tax: regulation of gene expression and disease. *Trends Microbiol.* **1**, 131–135.
- Zhuang, J., Mukherjee, S., Ron, Y. and Dougherty, J.P. (2006) High rate of genetic recombination in murine leukemia virus: implications for influencing proviral ploidy. *J. Virol.* **80**, 6706–6711.